Recurrence Plot Representation for Multivariate Time-series Analysis*

Dennys Mallqui, Msc.[1][0000-0003-2560-6407] and Ricardo A. S. Fernandes, PhD.[1]

Federal University of Sao Carlos, Sao Carlos 13565-905 SP, Brazil
dennys.mallqui@gmail.com, ricardo.asf@ufscar.br
https://www.linkedin.com/in/dennys-mallqui
http://ppgcc.dc.ufscar.br/pessoas/docentes-1/ricardo-fernandes

Abstract. The analysis of time-series is a productive field, which is applied in different areas such as finance, bio-medicine, neurology, among others. However, one of the main challenges is the identification of non-linear patterns. Thus, the apparent chaotic behavior of a time-series can mean the manifestation of a dynamic system. Often, these phenomena are recurrent, meaning that certain regions of their available state space are frequently visited along of time. For this reason, the use of recurrence plots (RPs) and Recurrent Quantification Analysis (RQA) are used to extract features of time series that allow their better understanding and facilitate prediction tasks (classification, regression and novelty detection). However, to successfully apply this transformation in the aforementioned tasks, it is necessary to obtain the best combination of three parameters: time lag (τ), embedding dimension (m) and recurrence rate (RR). In other studies to find these parameters it is necessary to apply the prediction process for each possible combination, which represents a high computational cost. We propose to use a measure that seeks to maximize the entropy with the lowest possible randomness to calculate $RP_{[\tau,m,RR]}$ before the application of the prediction. In this way, reduce the computational complexity, where we initially validate these claims using Bitcoin's multidimensional time-series, with results that surpass the accuracy of previous studies.

Keywords: Recurrence Plot · Recurrent Quantification Analysis · Time series · Chaos Theory · Multivariate classification · Bitcoin.

1 Introduction

The analysis of time-series for pattern recognition is an important field of research, which is applied in different fields of study such as financial data, biomedical signals, music mining and so on. Thus, the analysis tasks can be grouped into curve fitting or function approximation, forecasting, segmentation or classification and clustering as is detailed in [16]. One of the challenges encountered in the treatment of time-series is the identification of non-linear patterns, which

* This work has been supported/funded by KPiQa and Xertica companies.

are present in dynamic systems or stochastic processes [16, 12, 6]. Often, these phenomena are recurrent, meaning that certain regions of their available state space are frequently visited along of time [6].

Recurrence plots (RPs) and Recurrent Quantification Analysis (RQA) are tools of analysis of data that was initially introduced to understand the behavior of a dynamical system in phase space [27]. In recent studies [16, 12, 6, 17, 18, 7] these methods have been used with success to study patterns in the dynamics of a time-series system. According [9] it is possible to calculate the phase-space representation of a time-series X_t using a time lag or delay (τ) and embedding dimension (m) by computing the Delay Vectors (DVs) described in Equation 1.

$$DV(i) = [X_{i\text{-}m.\tau}, X_{i\text{-}(m\text{-}1).\tau}, ..., X_{i\text{-}\tau}] \quad (1)$$

In this sense, the recurrence plots can be calculated by considering the Equation 2, such as defined by [18]:

$$RP_{i,j} = \begin{cases} 1, & \|DV(i) - DV(j)\| \leq \varepsilon \\ 0, & otherwise \end{cases} \quad (2)$$

where ε is the distance metric threshold of two sampling points $DV(i)$, $DV(j)$. After that, the percentage of recurrence or recurrence rate (RR) is defined as a percentage of points where the distance function is less than ε between the total points. Thus, it is possible to represent the RP as three parameters $RP_{[\tau,m,RR]}$.

The application of RP presents high complexity when working with multivariate time-series, since the function of similarity between the embedded vectors can be complex to calculate. Also, an open problem is the identification of the most appropriate values of $RP_{[\tau,m,RR]}$ for time-series analysis (univariate and multivariate).

RP matrix exhibits global characteristic (typology) and local patterns (texture) as is mentioned in [16]. In the same study, the authors claims, that texture component is formed by single dots, diagonal lines as well as vertical and horizontal lines (using graphic information), while typology information which is characterized as homogeneous, periodic, drift and disrupted (using RQA measures). Thus, from the generated components, it is possible to analyze them to identify the inherent patterns in time series with chaotic behavior.

For instance, in [17] is used RP for image representation from multivariate time-series and using Convolutional Neural Networks (CNN) for identify patterns. In a similar study, [10] conduct an experiment for classification of DNA sequences where multivariate time-series are transformed in images by RP matrix and applied fractal dimensions (FD) for pattern detection. In other studies [2], [18], [9], the authors use RQA measures for novelty detection and drift signal identification. In addition, the RQA measures such as determinism (DET), laminarity (LAM) and DIV (1/Maximum length of diagonals) help to detection of change or classification. There are studies that propose the use of other measures to extract nonlinear patterns more efficiently in time-series. Thus, in [12] the authors used the Hotelling T^2 statistic for detection of dynamic transitions and compared to the use of RQA traditional measures. After that, in [7], it is proposed a Principal Singular Value Proportion (PSVP) measure to detect the complexity and periodicity of systems.

Although the use of texture analysis and typology have presented success in the above-mentioned studies, a serious problem still exists, which is the computational cost. This is because, for calculate RP matrix ($RP_{[\tau,m,RR]}$) is necessary to select adequately time lag or delay (τ), embedding dimension (m) and recurrence rate (RR). In [14], it is used minimum of the time-delayed mutual information as a reasonable value for τ and, in [8], it is selected the minimum embedding dimension m based on the false nearest neighbor algorithm. However, these methods were proposed for univariate time-series, with which the discussion of their obtaining in multivariate signals is still open. For this reason, studies in which predictions are made from the extracted RP and RQA features, iterate to find the optimal $RP_{[\tau,m,RR]}$ [6, 7, 16, 17], which implies a high computational cost.

Based on the context previously presented, this study seeks to focus on feature extraction for multivariate time-series using RP and RQA transformation, where a novel methodology will be sought to obtain the most appropriate values of delay (τ), embedding dimension (m) and recurrence rate (RR) for RP matrix calculation ($RP_{[\tau,m,RR]}$) without the need to iterate in the application of predictions, reducing the computational cost of this set of methods. For this, we propose to use a novel RQA measure that seeks to maximize the entropy with the lowest possible randomness inspired by [3]. Finally, once RP matrix is generated and the extraction of features is realized, we will proceed with the application of machine learning models for the identification of patterns on the transformed data and evaluate their contribution to tasks of classification using a Bitcoin time-series study conducted by the authors previously [23].

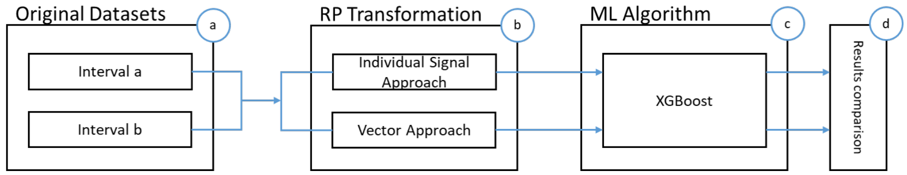
2 Experiment

An overview of the methodology is presented in Figure 2, where the following steps are shown: a) preparation and partition of the data set of the Bitcoin multivariate time-series; b) unsupervised features extraction using the RQA measures generated from RP matrix calculation using the proposed method; c) training a machine learning model using the transformed data and; d) results evaluation and comparison with other studies.

2.1 Data collected

A set of real-world multivariate time-series data was selected to validate the effectiveness of the proposed methodology. Thus, in this preliminary phase, it was selected the case study of Bitcoin trend prediction (daily exchange rate against US dollar). For traders or general users, the greatest challenge is the Bitcoin exchange rate volatility. Thus, as mentioned by [29] and [19], in comparison with the traditional currencies, the Bitcoin presents a volatility approximately

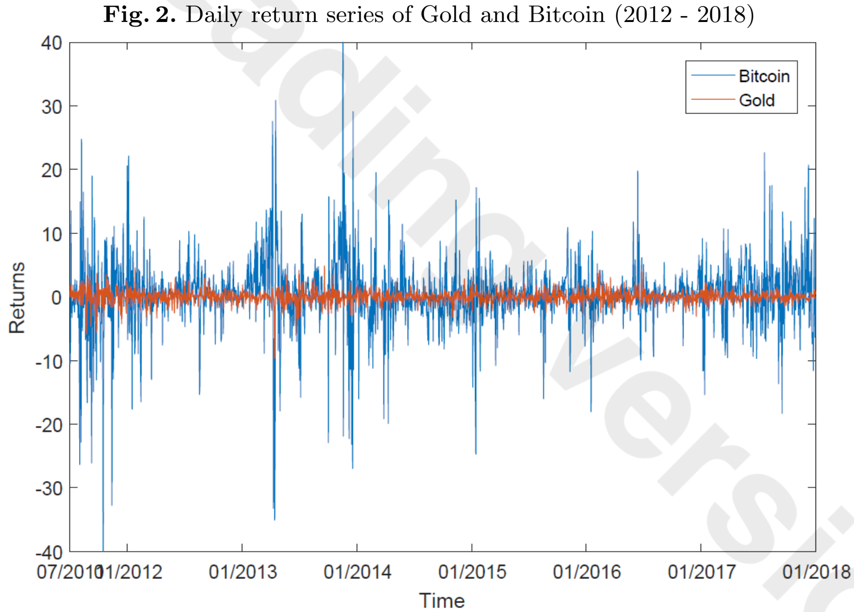
Fig. 1. Methodology Overview



Source: Personal collection.

twice greater referring to its value of exchange rate. For example, Figure **??** shows a comparison between the price volatility of the Bitcoin and the gold. However, as mentioned by [1], the author claims that the volatility of Bitcoin can not be a reason that invalidates it as a currency, but it is a motivation for users to seek solutions to reduce their risk [23].

Fig. 2. Daily return series of Gold and Bitcoin (2012 - 2018)



Source: [19].

The sorts of information can be categorized into *internal* (Bitcoin historical transaction data) and *external* (international economic factors or public recognition). Thus, in the data sets are considered Blockchain information (*internal*) that potentially includes Open, High, Low and Close (OHLC) Bitcoin prices, the volume of trades, total transaction fees, number of transactions, cost per transaction and average hash rate (measure about level of transaction peaks), as suggested by [25, 4]. As *external* information are considered international economic indicators where several possible indicators were analyzed, such as crude oil future prices, gold future prices, S&P500 future, NASDAQ future and DAX index [28]. In addition, it was considered public recognition data extracted from Google Trends and Wikipedia Searches, as used by [20, 13, 21]. Finally, using *internal* data was constructed technical indicators commonly used by traders, as proposed by [26].

2.2 Data partition

Such as used in [23], it is considered two data sets: (i) from 2013 to 2016 (named as *interval a* with 1066 instances), which considers 80%/20% for training/test; and (ii) from 2013 to 2017 (named as *interval b* with 1462 instances), which considers 75%/25% for training/test.

2.3 Attribute selection

It is considered the same attributes used on a previous study published by the authors [23]. Thus, in addition to technical indicators (Table 1) is included the most relevant attributes from Blockchain (Table 2 and 3), Economic indices (Table 4) and Social trends (Table 5) through information gain score.

Table 1. Attribute Selection - Technical Indicators for Interval a and b

Attribute	Details
OBV	On Balance Volume: $OBV_t = OBV_{t-1} + \theta \times V_t$
SMA_5	Simple Moving Average: $SMA_5 = (\sum_{i=1}^{5} C_{t-i+1}/5)$
$BIAS_6$	Average deviation: $BIAS_6 = (\frac{C_t - SMA_6}{SMA_6}) \times 100$
PSY_{12}	Psychological Line: $PSY_{12} = (A/12) \times 100$
ASY_5	Average of return (5 days): $ASY_5 = (\sum_{i=1}^{5} ASY_{t-i+1})/5$
ASY_4	Average of return (4 days): $ASY_4 = (\sum_{i=1}^{4} ASY_{t-i+1})/4$
ASY_3	Average of return (3 days): $ASY_3 = (\sum_{i=1}^{3} ASY_{t-i+1})/3$
ASY_2	Average of return (2 days): $ASY_2 = (\sum_{i=1}^{2} ASY_{t-i+1})/2$
ASY_1	Average of return (1 days): $ASY_1 = ASY_{t-1}$
Open. Price	Opening price in the same day of prediction

Source: Personal collection.

In these tables, V_t is the volume of trade of the Bitcoin at time t, θ is a step function, A is the number of rising days in the last n days, and WMA represent Weight Moving Average.

Finally, it was obtained the multivariate time-series with 19 attributes for trend prediction of daily Bitcoin exchange rate (classification task). In the following sections, it will be detailed two approaches that have been applied to extract the features of multivariate time-series.

Table 2. Attribute Selection - Blockchain data for Interval a

Attribute	Details
Transaction fees $D-2$	Voluntary fees paid by users to miners with time lag of 2 days
Hash rate average $D-2$	Average daily hash rate with time lag of 2 days
Minimum Price $D-5$	Minimum daily exchange rate with time lag of 5 days
Hash rate average $D-7$	Average daily hash rate with time lag of 7 days
Number of trx $30-Day\ WMA$	WMA of number of transactions in last 30 days

Source: Personal collection.

Table 3. Attribute Selection - Blockchain data for Interval b

Attribute	Details
Maximum Price $D-5$	Maximum daily exchange rate with time lag of 5 days
Minimum Price $D-5$	Minimum daily exchange rate with time lag of 5 days
Closing Price $D-5$	Closing daily exchange rate with time lag of 5 days
Volume of trades $D-5$	Volume of exchange transaction with time lag o 5 days
Hash rate avg $30-Day\ WMA$	WMA of daily hash rate average in last 30 days

Source: Personal collection.

Table 4. Attribute Selection - Economic indicators for Interval a and b

Attribute	Details
DAX index $30-Day\ WMA$	WMA of daily DAX index price in last 30 days

Source: Personal collection.

Table 5. Attribute Selection - Public recognition for Interval a and b

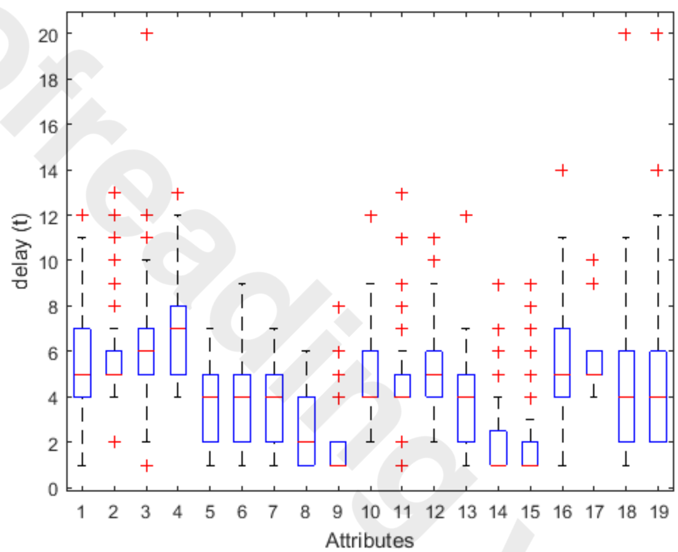
Attribute	Details
Google trends $W-4$	Google trend score with time lag of 4 weeks
Wikipedia trends $D-1$	Number of Wikipedia searches with time lag of 1 day
Wikipedia trends $30-Day\ WMA$	WMA of Wikipedia searches in last 30 days

Source: Personal collection.

2.4 RQA Transformation - Individual Signal Approach

In this stage, the approach to the treatment of multivariate time series is to individually analyze each of the signals that comprise it, generating for each one its corresponding RP and from it extracting RQA measurements that will finally be used for classification tasks. First, a time window is determined to create segments of each of the signals in isolation, each segment is calculated every 7 days (time step) considering a time window of 120 days backwards (window time). In each segment, the most suitable delay (τ) value is calculated from the method proposed by [14]. Then, Figures 3 and 4 show the distribution of the optimal τ values for each attribute, where it was select the average value represented by the 50th percentile in each case.

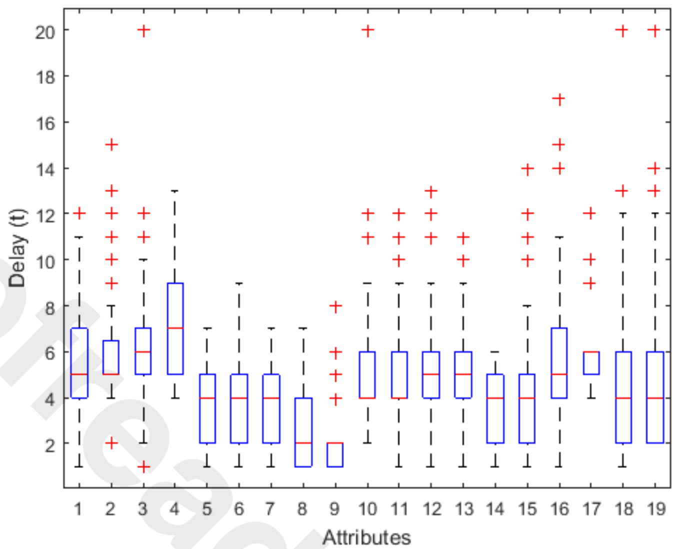
Fig. 3. Optimal Delay (τ) Selection – Interval a



Source: Personal collection.

As a next step, to obtain the most appropriate value of embedding dimension (m) and threshold (ε) for appropriate recurrence rate (RR), is proposed a novelty measure inspired by [3]. As mentioned by [24] the concept of entropy contains a concept of disorder, with which in consideration of the authors, it is sought to obtain the recurrent structure that presents a greater quantity of possible variations so that this allows to have more possibilities of identification of patterns. At this point, the proposed indicator is positively related to the amount of entropy that is detected by the RP matrix generated from a segment of the time-series. However, the exaggerated disorder can mean a randomness in the data which would invalidate any subsequent attempt to extract patterns in them. The indicator should be negatively related to the amount of randomness and, for this purpose, it will be used such as proposed by [7] (Principal Singular Value Proportion – PSVP). As mentioned by [3], in the arts, the most successful works of Abstract Expressionism show a random distribution of spray and splash pigment controlled by the artist's sense of visual order. Therefore, the

Fig. 4. Optimal Delay (τ) Selection – Interval b



Source: Personal collection.

Equation 3 tries to measure the balance between the amount of information or the disorder and order, therefore, we proposed the name of the measure as the Art Score (AS).

$$AS = \frac{ShannonEntropy}{PSVPscore} \tag{3}$$

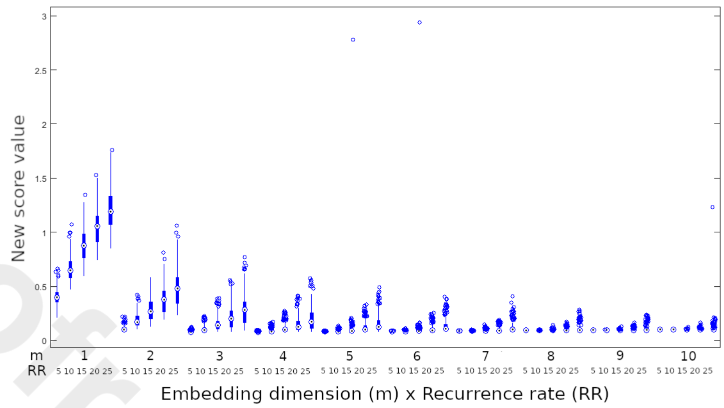
In addition, with the same segmentation to obtain the optimal delay, it is used to test combinations of embedding dimension (m: 1, 2, ..., 10) and recurrent rate (RR: 5%, 10%, ..., 25%) in order to obtain the highest AS value and, consequently, the optimal threshold (ε), as shown in Figures 5 and 6 for OBV attribute.

In Figure 5, the best result of AS in *interval a* is obtained by combination 5 ($m = 1$ and $RR = 25\%$). In a similar way, for *interval b*, the best result was obtained with the combination 5 ($m = 1$ and $RR = 25\%$). In summary, for each signal, it is tested multiple combinations of delay (τ), embedding dimension (m) and recurrence rate (RR) values as show in Table 6.

After that, with optimal τ, m and RR values, the RP matrix is calculated (best $RP_{[\tau,m,RR]}$) for each signal. Then, it is generated seven RQA measures for each one: recurrent rate (RR), determinism (DET), L average (Lavg), L maximum (Lmax), laminarity (LAM), transitivity (Trs) and PSVP.

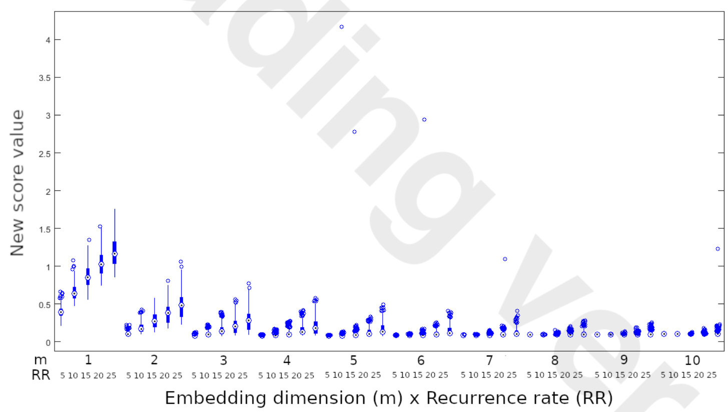
Because 133 variables are obtained in the individual processing of each of the 19 signals (19x7 features), which will be used for the algorithm proposed by [11], XGBoost. Thus, a selection is made based on the importance score obtained by

Fig. 5. Embedding Dimension (m) and Recurrent Rate (RR) – OBV (Interval a)



Source: Personal collection.

Fig. 6. Embedding Dimension (m) and Recurrent Rate (RR) – OBV (Interval b)



Source: Personal collection.

Table 6. Hyper-parameter combinations – Recurrence Matrix

Parameter	Values tested
Delay (τ)	1, 2, ..., 5
Embedding dimension (m)	1, 2, ..., 10
Recurrence rate (RR)	5%, 10%, ..., 25%

Source: Personal collection.

the model, using cross-validation method (10 folds). As a result, 38 features are selected for *interval a* and 45 features are selected for *interval b*.

2.5 RQA Transformation - Vector Approach

As proposed by [5], it was used the *vector recurrence* concept. In the Equation 2, $DV(i)$ and $DV(j)$ are d-dimensional vectors and the norm in R^d of vector difference is calculated. In addition, it is possible to consider any sort of norm in R^d like Euclidean or Mahalanobis distance. Thus, for this experiment, Euclidean norm is used with the same combinations presented in Table 6.

Once the optimal parameters for RP matrix ($RP_{[\tau,m,RR]}$) calculation have been determined, the corresponding matrix is generated and the following RQA measurements are extracted 14 features: recurrence rate (RR), determinism (DET), average diagonal length, maximum diagonal length, laminarity (LAM), transitivity (Trs), PSVP, diagonal entropy, average vertical length, maximum vertical length, average white vertical length, maximum white vertical length, global clustering and assortativity.

Finally, the information is compressed from 19 attributes to 14 attributes for *interval a* and *interval b*

3 Machine Learning Algorithm

Because for the individual transformation approach it was necessary to use XGBoost to obtain the ranking of the features and in order to compare both approaches, the algorithm proposed by [11] was used to generate the predictive models. In addition, the results of previous studies [22] were added using the original time-series (19 attributes), where the models used are based on Artificial Neural Networks (with prefix "*nn*") and Support Vector Machines (with prefix "*svm*").

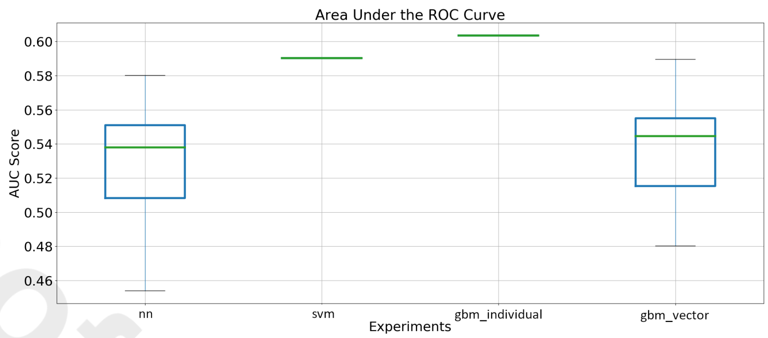
4 Results

In order to validate the prediction capacity of the unsupervised transformations, the Area Under the ROC Curve (AUC) metric is used as proposed by [15]. Thus, the values obtained by individual approach detailed in Section 2.4 (with prefix "*gbm_individual*") and vector approach mentioned in Section 2.5 (with prefix "*gbm_vector*") are compared.

In addition, all stochastic classification models were executed 50 times and, in Figures 7 and 8, are shown the average results obtained for *interval a* and *interval b*, respectively.

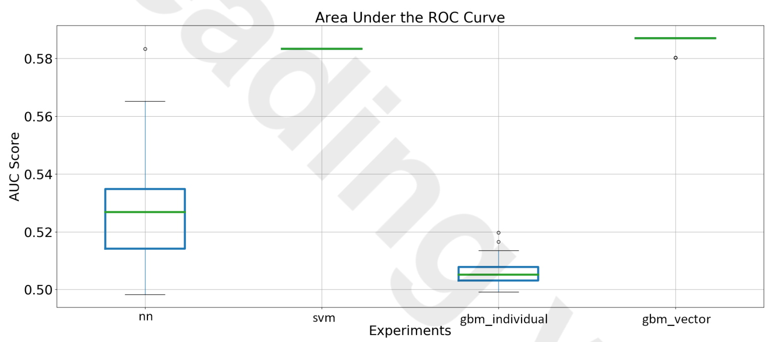
Finally, in *interval a* the best result is obtained by *individual approach* and in *interval b* the best score is presented by *vector approach*, where both proposed methods in present study surpass previous results.

Fig. 7. Classification Performance – AUC score (interval a)



Source: Personal collection.

Fig. 8. Classification Performance – AUC score (interval b)



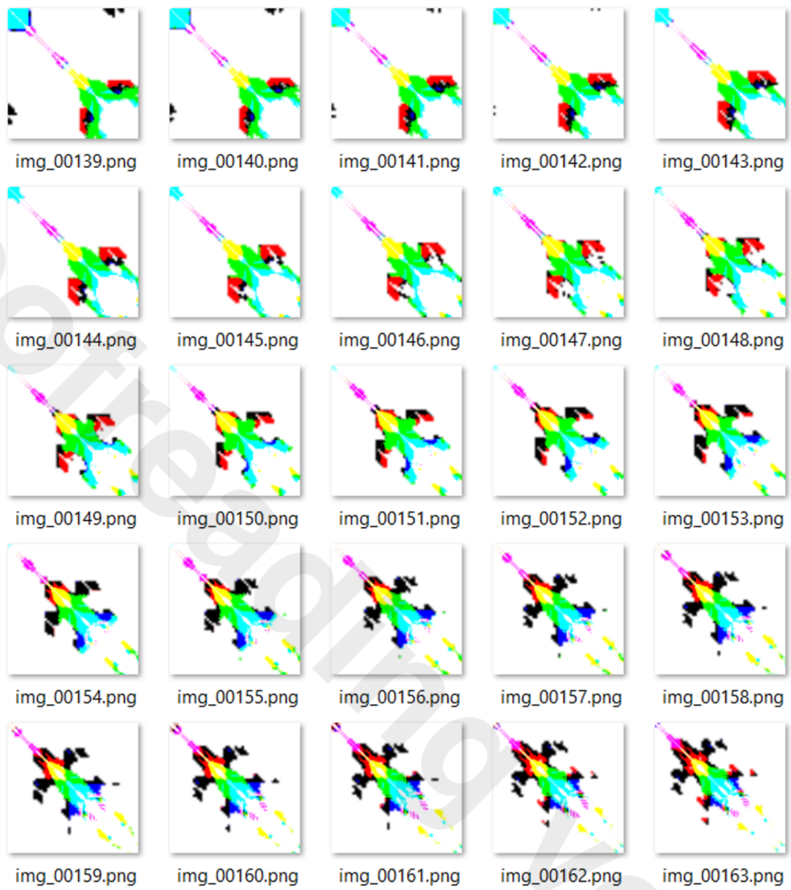
Source: Personal collection.

5 Conclusions and Future Research

The use of the proposed measure can improve the identification of patterns in time series, extracting characteristics in an unsupervised way, improving the results in the classification tasks and at the same time reducing the computational cost presented in previous studies. Therefore, we can infer that it is possible to use these transformations for future experiments that encompass tasks of identification of change of concept, regression and clustering.

Likewise, in addition to the transformations that use typology information (RQA measures), we consider that there is a great opportunity to take advantage of the texture features (RP images) due to its expressiveness to expose patterns in the signals as shown in Figure 9. Moreover, they could eventually improve the results obtained for the Bitcoin time-series and others in general as postulated by [17].

Fig. 9. RP Texture RGB – Slide windows of 120-90-60 days (interval b segment)



Source: Personal collection.

References

1. van Alstyne, M.: Why Bitcoin has value. Communications of the ACM **57**(5), 30–32 (2014)
2. Altmann, E.G., Kantz, H.: Recurrence time analysis, long-term correlations, and extreme events. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics **71**(5) (2005)
3. Arnheim, R.: Entropy and art: An essay on disorder and order. Univ of California Press (1974)
4. Balcilar, M., Bouri, E., Gupta, R., Roubaud, D.: Can volume predict bitcoin returns and volatility? A quantiles-based approach. Economic Modelling **64**, 74–81 (2017)
5. Bandt, C., Groth, A., Marwan, N., Romano, M.C., Thiel, M., Rosenblum, M., Kurths, J.: Analysis of bivariate coupling by means of recurrence. Understanding

Complex Systems **2008**, 153–182 (2008)
6. Beim Graben, P., Sellers, K.K., Fröhlich, F., Hutt, A.: Optimal estimation of recurrence structures from time series. Epl **114**(3) (2016)
7. Bian, S., Shang, P.: Recurrence quantity analysis based on singular value decomposition. Communications in Nonlinear Science and Numerical Simulation **46**, 1–13 (2017)
8. Cao, L.: Practical method for determining the minimum embedding dimension of a scalar time series. Physica D: Nonlinear Phenomena **110**(1-2), 43–50 (1997)
9. Carrión, A., Miralles, R.: New insights for testing linearity and complexity with surrogates: A recurrence plot approach. In: Springer Proceedings in Physics. vol. 180, pp. 91–112 (2016)
10. Cattani, C., Pierro, G., Cattani, C., Pierro, G.: On the Fractal Geometry of DNA by the Binary Image Analysis. Bulletin of Mathematical Biology **75**(9), 1544–1570 (2013)
11. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794. ACM (2016)
12. Chen, Y., Yang, H.: Multiscale recurrence analysis of long-term nonlinear and nonstationary time series. Chaos, Solitons and Fractals **45**(7), 978–987 (2012)
13. Ciaian, P., Rajcaniova, M., d'Artis Kancs: The economics of bitcoin price formation. Applied Economics **48**(19), 1799–1815 (2016)
14. Fraser, A.M., Swinney, H.L.: Independent coordinates for strange attractors from mutual information (1986)
15. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology **143**(1), 29–36 (1982)
16. Hatami, N., Gavet, Y., Debayle, J.: Bag of recurrence patterns representation for time-series classification. Pattern Analysis and Applications pp. 1–11 (2018)
17. Hatami, N., Gavet, Y., Debayle, J.: Classification of time-series images using deep convolutional neural networks (2018)
18. Hu, M.I.N., Zhou, S., Wei, J., Deng, Y., Qu, W.: Change-point Detection in Multivariate Time-series Data by Recurrence Plot 2 Related Works 3 Problem Formulation **13**, 592–599 (2014)
19. Klein, T., Pham Thu, H., Walther, T.: Bitcoin is not the new gold a comparison of volatility, correlation, and portfolio performance. International Review of Financial Analysis **59**, 105–116 (2018)
20. Kristoufek, L.: BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. Scientific Reports **3**, 1–7 (2013)
21. Li, X., Wang, C.A.: The technology and economic determinants of cryptocurrency exchange rates: The case of bitcoin. Decision Support Systems **95**, 49–60 (2017)
22. Mallqui, D.: Predicting the direction, maximum, minimum and closing price of daily/Intra-daily bitcoin exchange rate using batch and online machine learning techniques. Master's thesis, Universidade Federal de São Carlos (2018)
23. Mallqui, D., Fernandes, R.: Predicting the direction, maximum, minimum and closing prices of daily bitcoin exchange rate using machine learning techniques. Applied Soft Computing **75** (2018). https://doi.org/10.1016/j.asoc.2018.11.038
24. Marcolli, M.: Entropy and Art: the view beyond Arnheim (2015), `http://www.its.caltech.edu/~matilde/SlidesEntropyArt.pdf`
25. Mcnally, S.: Predicting the price of Bitcoin using Machine Learning. Ph.D. thesis, National College of Ireland (2016)
26. Qiu, M., Song, Y.: Predicting the direction of stock market index movement using an optimized artificial neural network model. PLoS ONE **11**(5), 1–11 (2016)

27. Romano, M.C., Thiel, M., Kurths, J., Von Bloh, W.: Multivariate recurrence plots. Physics Letters, Section A: General, Atomic and Solid State Physics **330**(3-4), 214–223 (2004)
28. Vassiliadis, S., Papadopoulos, P., Rangoussi, M., Konieczny, T., Gralewski, J.: Bitcoin Value Analysis Based On Cross-Correlations. Journal of Internet Banking and Commerce **22**(S7), 1 (2017)
29. Yermack, D.: Is Bitcoin a real currency? An economic appraisal. National Bureau of Economic Research (NBER) Working Papers **19747** (2013)